CORRELATION AND MACHINE CALCULATION

By H. A. Wallace and George W. Snedecor

IOWA STATE COLLEGE OF AGRICULTURE AND MECHANIC ARTS

DIVISION OF INDUSTRIAL SCIENCE

DEPARTMENT OF MATHEMATICS

January, 1925

AMES, IOWA

IOWA STATE COLLEGE
OF AGRICULTURE AND MECHANIC ARTS
OFFICIAL PUBLICATION

Vol. 23 January 28, 1925 No. 35

CORRELATION AND MACHINE CALCULATION

by

H. A. WALLACE
Editor "Wallaces' Farmer"

and

GEORGE W. SNEDECOR
Associate Professor of Mathematics
Iowa State College

DIVISION OF INDUSTRIAL SCIENCE

Department of Mathematics

AMES, IOWA

Published weekly by Iowa State College. Entered as second-class matter and accepted for mailing at special rate of postage, Act. Aug. 24, 1912, authorized Apr. 1, 1920.

CONTENTS

Copies of this bulletin may be obtained from the Iowa State College Book Store, Ames, Iowa, at 50c each.

CORRELATION
AND
MACHINE CALCULATION

The rapid extension during recent years of the ideas of simple correlation has imposed their use upon many scientists not trained in the mathematical theory underlying them. The present trend in all biological sciences, as well as in economics and psychology, is still further to extend the use of correlation, broadening its scope to include the associations among more than two variables. One object of this bulletin is to present in simple, untechnical language some explanation of the meaning and uses of the various correlation coefficients, simple, partial and multiple.

The second and principal object of the bulletin is to set forth explicit directions for the use of the usual commercial forms of calculating machines, either key-driven, such as the Comptometer and Burroughs Calculator, or crank driven, such as the Monroe or Marchant, in finding correlation coefficients or related constants. According to the usual procedure, where the arithmetic is done mentally, the use of the correlation table, or double entry table, is almost indispensable. The advent and prevalent use of calculating machines, however, make practicable a return to simpler and more direct methods of reckoning. These machines are admirably adapted to the calculation of all the correlation constants with speed and precision.

For extensive data where the number of observations runs into the thousands, punched cards should be used with sorting and tabulating machines, such as the Hollerith machines. The average research worker, however, who is dealing with less than 500 cases, will probably find the methods herein set forth well adapted to his use.

For the benefit of those readers who are not familiar with the ideas of simple correlation between two variables, we shall present them very briefly in the following paragraphs.

PART I. SIMPLE CORRELATION.

A simple correlation coefficient, r, between two variables is a measure of the degree to which they tend to be associated or to move together. If they should move in the same direction, keeping perfect step all the way, r is so designed as to take the value,

+ 1; if on the contrary they should move in exactly opposite directions, but at the same proportional rates, r then assumes the value, — 1. In actual statistical work, such perfect associations do not occur, and r will usually be found to lie somewhere between —.95 and +.95. As an example, if we take the size of the corn crop and the price of corn per bushel year by year for a number of years back, we find that r = — .78. This is the numerical measure, furnished by the methods of correlation, of the very real tendency of large corn crops to be associated with low prices, and *vice versa*.

It is interesting to observe why the letter r is universally used to designate a correlation coefficient. Sir Francis Galton, who first used the idea of correlation, as here presented, back in the early 1880's was working on the problem of the degree to which children inherited height from their parents. He looked upon the tendency of children to resemble their parents only partly while partly reverting to racial characteristics, as a "regression" of the inherited characteristics upon those of the parents. Originally, therefore, r stood for regression or reversion.

TABLE 1. CORN YIELD AND LAND VALUE

Observation number	County	Average Corn Yield in bushels per acre, 1910-1919	Average land value per acre, Jan. 1, 1920
		A	X
1	Allamakee	40	$ 87
2	Bremer	36	133
3	Butler	34	174
4	Calhoun	41	285
5	Carroll	39	263
6	Cherokee	42	274
7	Dallas	40	235
8	Davis	31	104
9	Fayette	36	141
10	Fremont	34	208
11	Howard	30	115
12	Ida	40	271
13	Jefferson	37	163
14	Johnson	41	193
15	Kossuth	38	203
16	Lyon	38	279
17	Madison	34	179
18	Marshall	45	244
19	Monona	34	165
20	Pocahontas	40	257
21	Polk	41	252
22	Story	42	280
23	Wapello	35	167
24	Warren	33	168
25	Winneshiek	36	115

In fundamental research in the biological, agricultural, economic and educational fields, correlation is often of the greatest service in demonstrating the value of hypotheses already tentatively accepted, and in suggesting new hypotheses for verification.

Further explanation and discussion will be made in connection with the data of Table 1, consisting of paired observations on the corn yield in bushels per acre (average for years 1910-1919), and land value per acre (Jan. 1, 1920) in 25 Iowa counties.

The numbers giving corn yield will be designated by the symbol A; those giving land value, by X. The value X is to be thought of as "dependent" upon the yield, A. Hence X is to be considered as the "criterion" or "dependent variable," while A is called the "independent variable".

We shall first explain in detail the procedure of calculation, summarizing the results and formulas in Table 2, and shall then discuss the meaning of the results obtained.

First. Add each column of observed values on the machine, designating the sum of the A-values by ΣA (Σ is the Greek equivalent of S. The symbol ΣA is read either "Sigma A" or better "sum of the A's.") and the sum of the X-values by ΣX; that is

$$\Sigma A = 40 + 36 + 34 + \text{etc.}, = 937$$
$$\Sigma X = 87 + 133 + 174 + \text{etc.}, = 4{,}955$$

Second. Using the machine, divide each sum by the number of observations, 25, the results being the arithmetic means (averages) of the variables. In the formulas, the number of observations is denoted by n. If we designate the means by M_A and M_X respectively, we then have,

$$M_A = \frac{\Sigma A}{n} = \frac{937}{25} = 37.48 \text{ bu. per acre}$$

$$M_X = \frac{\Sigma X}{n} = \frac{4955}{25} = \$198.20 \text{ per acre}$$

Third. Calculate the sum of the squares of the individual A's, thus,

$$\Sigma A^2 = (40)^2 + (36)^2 + (34)^2 + \text{etc.}, = 35{,}461$$

The individual values should be squared on the machine (40×40), (36×36), etc., and the sum carried through the entire twenty-five operations *without clearing the machine*. Thus, no record is made of the individual squares, but only of their sum. Similarly,

$$\Sigma X^2 = (87)^2 + (133)^2 + (174)^2 + \text{etc.}, = 1{,}075{,}817$$

Fourth. Calculate the sum of the products of the pairs of values:

$$\Sigma AX = (40 \times 87) + (36 \times 133) + (34 \times 174) + \text{etc.}, = 189{,}533$$

These products are also carried *without clearing the machine,* so that the total in the machine at the end of the twenty-five multiplications is the required "product-moment".

A word of explanation should be interpolated as to the number of decimal places carried in this illustrative example. Since $M_A = 37.48$ bu. $\pm$ .50 bu., only the first decimal place has statistical significance. $M_X = \$198.20 \pm \$8.26$, so that the number of cents need not be carried at all. (The reliability of r will be discussed later.) The arithmetical work has usually been carried to four decimal places, however, so that the reader may verify the operations without confusion. This is merely for convenience and uniformity and is not intended to denote statistical significance.

The results of the four operations just described (Lines 1, 2 and 3, Table 2) constitute the data for the calculation of the correlation coefficient between A and X. The formula used is a form of the ordinary product-moment formula, as follows:

$$r = \frac{\Sigma AX - (\Sigma A)M_X}{\sqrt{\Sigma A^2 - (\Sigma A)M_A} \times \sqrt{\Sigma X^2 - (\Sigma X)M_X}}$$

This may be read in words, "The correlation coefficient between X and A is given by a fraction whose numerator is the difference between ΣAX and the product of ΣA by M_X. The denominator is the product of two square roots, the first being the square root of the difference between ΣA^2 and the product of ΣA by M_A; the second, the square root of the difference between ΣX^2 and the product of ΣX by M_X."

This is almost the same form for r as that given in Rietz's "Handbook of Mathematical Statistics", page 122. It is in no sense an approximation, but is derived by ordinary algebraic processes from the more usual forms.

The calculation is now completed in the following steps, beginning (where we left off) with the

Fifth. Calculate the three products,

$$(\Sigma A)M_A = 937 \times 37.48 = 35{,}119$$
$$(\Sigma X)M_X = 4{,}955 \times 198.20 = 982{,}081$$
$$(\Sigma A)M_X = 937 \times 198.20 = 185{,}713$$

Enter these results as indicated in Line 4, Table 2.

TABLE 2. SUMMARY OF FORMULAS AND CALCULATIONS

1.	$\Sigma A = 937$	$\Sigma X = 4{,}955$	$n = 25$
2.	$M_A = 37.48$ bu.	$M_X = \$198.20$	
3.	$\Sigma A^2 = 35{,}461$	$\Sigma X^2 = 1{,}075{,}817$	$\Sigma AX = 189{,}533$
4.	$(\Sigma A)M_A = 35{,}119$	$(\Sigma X)M_X = 982{,}081$	$(\Sigma A)M_X = 185{,}713$
5.	$\Sigma A^2 - (\Sigma A)M_A = 342$	$\Sigma X^2 - (\Sigma X)M_X = 93{,}736$	$\Sigma AX - (\Sigma A)M_X = 3{,}820$
6.	$\sqrt{\Sigma A^2 - (\Sigma A)M_A} = 18.49$	$\sigma_A = 3.70$ bu.	$\sqrt{\Sigma X^2 - (\Sigma X)M_X} = 306.16$
7.		$\sigma_X = \$61.23$	$r = .6747$

Sixth. Subtract the numbers in Line 4 from those in Line 3 as indicated in Table 2.

Seventh. Extract the square roots of the first two results just obtained; thus,

$$\sqrt{\Sigma A^2 - (\Sigma A)M_A} = \sqrt{342} = 18.49$$
$$\sqrt{\Sigma X^2 - (\Sigma X)M_X} = \sqrt{93{,}736} = 306.16$$

Square roots may be calculated on the machine or by the usual arithmetic methods, but a table of squares and square roots (such as Barlow's) gives the results much more rapidly.

Eighth. Multiply the two results just obtained to get the denominator of the fraction in the formula for r;

$$\sqrt{\Sigma A^2 - (\Sigma A)M_A} \times \sqrt{\Sigma X^2 - (\Sigma X)M_X} = 18.49 \times 306.16 = 5{,}660.9$$

Ninth. To obtain r, divide as indicated in the formula;

$$r = \frac{3{,}820}{5{,}660.9} = .6747$$

thus completing the calculation of r. However, for later use we shall add to Table 2 the following:

Tenth. Compute the standard deviation of the A's thus,

$$\sigma_A = \frac{\sqrt{\Sigma A^2 - (\Sigma A)M_A}}{\sqrt{n}} = \frac{18.49}{5} = 3.70 \text{ bu. per acre}$$

Also, for the standard deviation of the X's,

$$\sigma_X = \frac{\sqrt{\Sigma X^2 - (\Sigma X)M_X}}{\sqrt{n}} = \frac{306.16}{5} = \$61.23 \text{ per acre}$$

For the benefit of the novice in the use of a calculating machine, it is suggested that sub-totals be recorded frequently when a long column is being added, especially if multiplications are being done at the same time. This helps in checking the results. An experienced operator will check his work 49 times out of 50 the second time over. The following gives some idea of the speed that may be maintained by fairly proficient operators in carrying through the various operations in the problem just completed:

APPROXIMATE TIME OF CALCULATIONS

	Time of operations in seconds					Total time in minutes including entries of results	Total time in minutes of checking calculations
	ΣA	ΣX	ΣA^2	ΣX^2	ΣAX		
Key Driven Machine	15	20	45	100	90	7½	5
Crank Driven Machine	45	55	115	145	140	12	10

It should be distinctly understood that these figures are given merely for the guidance of the novice, and have little or no bearing on the relative merits of key and crank driven machines. Each type of machine has peculiar advantages, and the type to be used in any given office depends upon many circumstances besides the speed attained in the calculation of this particular problem.

In the following sections will be set forth the meaning and uses of simple correlation coefficients, using the one just calculated as an illustration. The beginner is warned not to attempt a too literal interpretation. Although perfect correlation is measured by 1.00, the $r = .67$ (we shall carry only the first two decimal places in this discussion) cannot be thought of as a percent. There is no absolute scale on which we can say that one correlation is high and another low.

RELIABILITY OF THE CORRELATION COEFFICIENT

As is the case with all statistical constants, the reliability of a correlation coefficient is indicated by the smallness of its standard deviation. Denoting the standard deviation of r by the symbol, σ_r, the formula is,

$$\sigma_r = \frac{1 - r^2}{\sqrt{n}}$$

In our example

$$\sigma_r = \frac{1 - (.67)^2}{\sqrt{25}} = .11$$

To interpret this in connection with our r of .67, we first calculate the range from $(.67 - .11)$ to $(.67 + .11)$; that is, the range from .56 to .78. We then say (from theoretical considerations) that if these data were collected over and over from similarly located counties the chances are that about 68% of the resulting r's would lie between .56 and .78. Of the other 32%, about half would lie below .56 and the remainder above .78.

If the reader is more familiar with the idea of "probable deviation" (probable error) then he may use the formula,

$$E_r = .6745 \frac{1 - r^2}{\sqrt{n}},$$

which gives a probable deviation of .07 in our example. The corresponding range is now from .60 to .74 $(.67 \pm .07)$ and the interpretation is that in future experiments similarly conducted we may expect about 50% of the resulting r's to lie within this range.

It is now evident that only the first two decimal places in r have statistical significance. As will appear later, the arithmetical operations are standardized by carrying the calculations to four places of decimals, but this is done merely for convenience in verifying the results. For an excellent short statement as to the number of significant figures, see Truman L. Kelley, "How Many Figures are Significant?" in Science, Vol. LX, No. 1562, page 524, Dec. 5, 1924.

It is perhaps simpler to calculate a range within which all r's would be likely to lie. While certainty is unattainable, we may say that a range of twice the standard devation will usually contain above 95% of similarly obtained r's, while a range of three times the standard deviation will probably contain more than 99% of them. The first of these ranges is sufficient for ordinary practical work, while the second would be accepted for most scientific work. In our example, $2\sigma_r = 2 \times .11 = .22$. We may therefore reasonably expect 95% of similarly obtained r's to lie between .45 and .89 $(.67 \pm .22)$. Since $3\sigma_r = .33$, the range $.67 \pm .33$ (from .34 to 1.00) will probably contain all similarly calculated r's.

It is now easily understood why reliability is measured by the

smallness of the standard deviation. The smaller the range necessary to include 95% of all similarly calculated r's, the more likely it is that such r's will closely approximate the one already obtained. Study of the formula for σ_r will reveal that two elements enter into the determination of its smallness; first, the largeness of r itself, and second, the largeness of the number of observations. If, for example, our r had been .77 instead of .67, its standard deviation would have been

$$\sigma_r = \frac{1 - (.77)^2}{5} = .08$$

and the smaller range from .61 to .93 ($.77 \pm 2 \times .08$) would be likely to embrace 95% of such r's. On the other hand, if the original $r = .67$ had been obtained from 100 observations instead of 25, the corresponding standard deviation would have been

$$\sigma_r = \frac{1 - (.67)^2}{\sqrt{100}} = .055$$

just half of the actual value in the given example. The correspondingly smaller range from .56 to .78 ($.67 \pm 2 \times .055$) would then contain 95% of such r's.

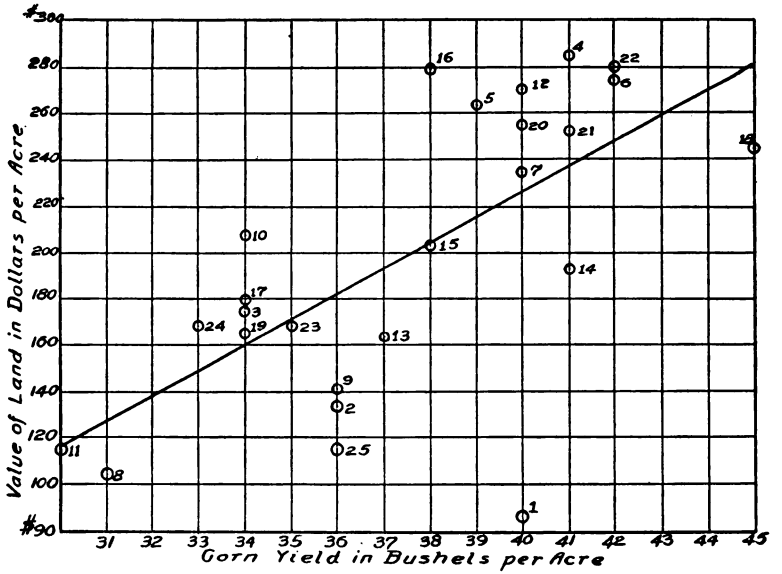
The student who will experiment with the formula, testing for reliability r's of various sizes and depending upon different numbers of observations, will soon gain a real appreciation of the way in which their reliability depends upon these elements.

THE REGRESSION EQUATION

The most practical use of r is in the calculation of the equation of the "regression line", whose meaning and use will now be discussed. Fig. 1 shows the familiar dot diagram, or scatter diagram, of the data of Table 1. One dot, properly located, represents each pair of values in the table. The fact that A and X are correlated is shown qualitatively on this diagram by the distribution of the dots in a band and not merely at random. The regression line shows the trend of this band of dots. It represents the best average position of the dots that statistical study is able to furnish. The line is plotted on the diagram and the estimated land values of Table 3 are calculated by means of the "regression equation" (in which the symbol $\overline{X}$ is read "estimated value of X"),

$$\overline{X} = M_X + r \times \frac{\sigma_X}{\sigma_A} (A - M_A)$$

FIGURE 1. THE REGRESSION LINE



If we substitute in this formula the values computed in our example, we have,

$$\overline{X} = 198.20 + .6747 \times \frac{61.23}{3.70} (A - 37.48),$$

or, performing indicated arithmetical operations,

$$\overline{X} = 11.17A - 220.45$$

This means that for a particular corn yield, say A = 38 bu. per acre, the corresponding estimated land value will be $11.17 \times 38 - 220.45 = \$204.01$ per acre.

Continuing the calculations as above of estimated values from actual corn yields, we have the values appearing in Table 3. (The number of cents is not recorded as it has no statistical significance.)

There are two counties, Kossuth and Lyon, whose average corn yield is 38 bu. per acre. It will be observed that the corresponding estimated land value of approximately $204 per acre agrees closely with the actual land value of Kossuth county, but is $75 too low for Lyon county. The estimated land value is a kind of average value (not the arithmetic mean) corresponding to this particular figure for corn yield, but taking account of the peculiarities not only of these two counties, but also of all the

TABLE 3. LAND VALUE ESTIMATED FROM CORN YIELD

Observation	Actual Average land value per acre, Jan. 1, 1920	Estimated land value per acre	Errors of Estimate
1. Allamakee	$ 87	$226	—139
2. Bremer	133	182	— 49
3. Butler	174	159	15
4. Calhoun	285	238	47
5. Carroll	263	215	48
6. Cherokee	274	249	25
7. Dallas	235	226	9
8. Davis	104	126	— 22
9. Fayette	141	182	— 41
10. Fremont	208	159	49
11. Howard	115	115	..
12. Ida	271	226	45
13. Jefferson	163	193	— 30
14. Johnson	193	238	— 45
15. Kossuth	203	204	— 1
16. Lyon	279	204	75
17. Madison	179	159	20
18. Marshall	244	282	— 38
19. Monona	165	159	6
20. Pocahontas	257	226	31
21. Polk	252	238	14
22. Story	280	249	31
23. Wapello	167	170	— 3
24. Warren	168	148	20
25. Winneshiek	115	182	— 67

twenty-five counties. To the student of statistics, the case of Lyon county is the more interesting and important. He immediately asks, "What peculiarity has this county that makes its land value diverge so greatly from the estimated or average value?" This is exactly the question whose answer must be found in the later chapters on multiple correlation. Corn yield is only one of the many characteristics entering into the determination of land value. An examination of Table 6 will show that whereas Lyon county has close to the average corn yield, its percentages of farm land in corn and small grain, and its number of brood sows per thousand acres are all much higher than average. Multiple correlation is a scheme for taking into account associations of all these elements with land value.

We now come to the real problem in any statistical study—how to interpret the results. As indicated above, one of the most fruitful sources of information is the study of the cases in which the estimated values diverge most widely from the actual land values. It should be noticed that these "errors of estimate" are positive or negative according as the estimated value falls short of the actual value, or exceeds it. We have just discussed the

largest of the positive errors of estimated land value, that of Lyon county; let us now study the largest of the negative errors. Allamakee county has a land value of $139 less than that predicted on the basis of corn yield alone. Referring again to Table 6, we find that although this county has a high corn yield per acre, its percentages of land in corn and small grains are among the lowest of the 25 counties. It is obvious that we shall have to include one or both of these elements in our study.

In order to help the reader to get a correct concept of the kind of averages these estimated values are, we call his attention to the following facts:

1. The sum of the positive errors of estimate (actual values minus estimated values) is equal to the sum of the negative errors; that is, the algebraic sum of all the errors is zero. Hence, while any one of these estimated (average) values may deviate considerably from the actual, they are so adjusted that the algebraic sum of all such deviations is the least possible.

2. The sum of the squares of these errors of estimate is less than it would be if any other linear regression equation had been used. Since the "standard error of estimate" (to be explained later) is calculated directly from such sum, it follows that the standard error of estimate is less than any other root-mean-square average of such errors of estimate.

All this means that the regression line is drawn through the dots in such a way that the algebraic sum of the vertical distances of all the dots from the line is zero, and the sum of the squares of such distances is a minimum.

An interesting interpretation of the meaning of r can now be introduced. It is not only the correlation coefficient between corn yield and land value, but is also the correlation coefficient between actual land value and estimated land value. That is, if r were calculated for the two columns of land values in Table 3, its value would be .67. It is therefore in a very definite sense a direct measure of our success in estimation.

It should be clearly understood that there are two regression lines in simple correlation. The one just discussed is known as the regression of land value on corn yield; that is, the regression of X on A. If we should wish to estimate average corn yield from given land values we should have to use the formula for regression of A on X, as follows:

$$\overline{A} = M_A + r \times \frac{\sigma_A}{\sigma_X} (X - M_X)$$

Using the values computed in our example, this reduces to

$$\overline{A} = .0408X + 29.4$$

This represents a different line from that shown in Fig. 1; and any particular land value together with the estimated yield as computed from this last formula constitute a different pair of values from those found in Table 3. Only in the case of perfect correlation ($r = 1$) would there be perfect agreement between the pairs of values calculated from the two regression formulas.

THE STANDARD ERROR OF ESTIMATE

The second practical use of r is to enable us to calculate easily an average of the differences between the actual and estimated land values; that is, an average of the errors of estimate given in the last column of Table 3. As already indicated, the mean of these errors of estimate is zero, because their algebraic sum is zero. The average which is generally used is the root-mean-square average known as the "standard error of estimate". It might be obtained by squaring each of the numbers in this last column of Table 3, adding such squares, dividing the sum by 25 (the number of observations) and extracting the square root of the quotient. Practically, however, the same result (which we shall designate by $\sigma_{X \cdot A}$) is obtained by the use of the formula

$$\sigma_{X \cdot A} = \sigma_X \sqrt{1 - r^2}$$

Substituting our values, we find that the standard error of estimated land values is

$$\sigma_{X \cdot A} = 61.23 \sqrt{1 - (.6747)^2} = 61.23 \times .738 = \$45.19 \text{ per acre}$$

It is to be observed that this standard error of estimate is 73.8% of the standard deviation, σ_X. In other words, the standard deviation of predicted values from actual values is only 73.8% of the standard deviation of actual values from their mean.

Since this standard error of estimate is the standard deviation of the differences between actual and estimated land values, it has the usual interpretation of standard deviations (see page 11); that is, about 68% of the errors lie in the range from $-$ \$45.19 to $+$ \$45.19. Furthermore, approximately 95% of the errors are expected to lie in the range from $-$ \$90.38 to $+$ \$90.38; and usually all the errors are included in the range from $-$ \$135.57 to $+$ \$135.57. An examination of the actual errors will convince the reader of the close agreement of these theoretically computed ranges with the facts.

A somewhat different interpretation of the standard error of estimate, $\sigma_{X \cdot A}$, may make its meaning and importance clearer. If we are asked to estimate the average land value of one of our 25 counties knowing nothing of its average corn yield, we shall have to be satisfied with the following answer; it is more likely to be worth about \$198.20 per acre (the mean value, M_X) than

any other amount, and the standard error of all such estimated values is \$61.23 per acre (the standard deviation, σ_X). If, however, knowing that the r between land value and corn yield is .67, we are given the additional information that the corn yield of a particular county is 38 bu. per acre, then we are able to better our estimate in two respects. We are able to say from the regression formula that the land value is more likely to be around \$204 per acre than any other value, a much better estimate than before; and we are also able to say that the standard error of such estimates is now only \$45.19 ($\sigma_{X \cdot A}$), or only 73.8% of σ_X, thus indicating greater reliability in the estimations.

The question arises—how much better is an r of .6 than one of .4? What does the relative size of the r's mean? To answer this, we say that if $r = .6$, then

$$\sigma_{X \cdot A} = \sigma_X \sqrt{1 - (.6)^2} = .8\ \sigma_X \text{ or } 80\% \text{ of } \sigma_X,$$

while if $r = .4$, then

$$\sigma_{X \cdot A} = \sigma_X \sqrt{1 - (.4)^2} = .917\ \sigma_X \text{ or } 91.7\% \text{ of } \sigma_X.$$

Thus, an r of .6 reduces the standard deviation of estimated values by 20%, whereas an r of .4 reduces it by only 8.3%. The following table gives the percentages by which different r's reduce the standard deviations of estimated values:

TABLE 4. REDUCTION OF STANDARD DEVIATION

r	Pct. Reduction of Standard Deviation	r	Pct. Reduction of Standard Deviation	r	Pct. Reduction of Standard Deviation
.05	.1%	.50	13.4%	.92	60.8%
.10	.5%	.55	16.5%	.94	65.9%
.15	1.1%	.60	20.0%	.95	68.8%
.20	2.0%	.65	24.0%	.96	72.0%
.25	3.2%	.70	28.6%	.97	75.7%
.30	4.6%	.75	33.9%	.98	81.0%
.35	6.3%	.80	40.0%	.99	85.9%
.40	8.3%	.85	47.3%	.999	95.5%
.45	10.7%	.90	56.4%	1.000 or prediction perfect	100.0%

From this table it is possible to say that for estimating purposes an r of .8 reduces the standard deviation twice as much as does an r of .6. Correlations of less than .4 are evidently practically worthless for estimating purposes because they reduce the standard deviation of estimated values by less than 9%. This table also shows very strikingly why it is that correlation coefficients cannot be interpreted as percentages.

Beginners almost invariably attach more significance than they

should to low correlations and less than they should to higher correlations. In working with multiple correlations they sometimes find that the inclusion of two more variables will raise the correlation from .90 to .94. They hesitate to go to the labor of the extra calculation because they think that a gain of .04 is insignificant. This is a mistake. The table indicates that an r of .90 reduces the standard deviation by 56.4%, whereas an r of .94 reduces it by 65.9%. An additional reduction of 9.5% in the standard deviation of estimated values is tremendously worth while. In this sense, an r of .94 excels an r of .9 by more than an r of .4 excels one of .05.

In concluding this part on simple correlation it should be observed that all of the interpretations and conclusions are valid only in so far as the band of dots in the scatter diagram approximates rectilinearity. Any pronounced curve in this band indicates a curvilinear regression line. The validity of the methods herein described decreases as the curvilinearity of the regression line increases. This applies to the following pages also. It should be emphasized that scatter diagrams should always be plotted before proceeding with any calculations of correlation. Should there be a pronounced curvilinearity of regression it may or may not be practicable to divide the entire range into two or more sections and treat them separately. Of course, the results obtained will be valid only within the particular range or ranges treated. Another recourse is to fit a curve to the regression as an empirical formula. The resulting regression equation is used for estimation purposes exactly like those herein described. Some interesting work is being done on this problem by Dr. F. C. Mills and Mr. Mordecai Ezekiel, their results appearing in some of the current numbers of the Journal of the American Statistical Society.

PART II. MULTIPLE CORRELATION—THREE VARIABLES.

Whenever we dig thoroughly into any problem, we generally find it necessary to study a whole net-work of relationships. In the case of land values, we soon found other variables besides corn yield affecting land value. In the present chapter we shall consider the solution of the problem of one additional variable introduced—percentage of farm land in small grains. The values of this new variable, given in Table 6, will be designated by the letter B.

First, of course, we shall have to calculate in the same way as before the following constants for the new series of observed values:

$$\Sigma B = 488,\ M_B = 19.52\%,\ \Sigma B^2 = 10{,}418,\ \Sigma B^2 - (\Sigma B)M_B = 892,$$
$$\sqrt{\Sigma B^2 - (\Sigma B)M_B} = 29.87,\ \sigma_B = 5.97\%.$$

Next, in order to find the new simple correlations required, we shall need

$$\Sigma BX = 104{,}064, \qquad \Sigma AB = 18{,}519,$$
$$\Sigma BX - (\Sigma B)M_X = 7{,}342 \qquad \Sigma AB - (\Sigma A)M_B = 229.$$

Now, the two correlation coefficients are found exactly as in Part 1, a simple interchange of letters giving the new formulas.

In order to distinguish the three r's, we shall add as subscripts the letters corresponding to the variables correlated; thus r_{AX} is the one already found—the correlation coefficient between corn yield A, and land value X. Denote by r_{AB} the correlation coefficient between the percentage of land in small grain B, and corn yield A; and by r_{BX} the correlation coefficient between percentage of land in small grain B and the land value X. The order in which the subscripts are written has no significance; that is, r_{XA} is the same coefficient as r_{AX}, etc. The new calculations result in, $r_{AB} = .4146$ and $r_{BX} = .8029$. These two new r's together with the one formerly found ($r_{AX} = .6747$) constitute the necessary data for the calculation of the multiple correlation coefficient which is always denoted by R, and the multiple regression equation.

From this point on, the method of procedure differs radically from that of Part I. In the first place we shall have to introduce two new quantities, the "partial regression coefficients". Their use will appear presently. They are denoted by the symbols β_{XA} (β is the Greek letter for b, and the symbol is read "beta X A") and β_{XB}. The two β's are found by solving a pair of simultaneous equations known as "normal equations". (See Kelley: "Statistical Method", p. 282.) In symbolical form these two equations are written

$$\beta_{XA} + r_{AB}\beta_{XB} = r_{AX},$$
$$r_{AB}\beta_{XA} + \beta_{XB} = r_{BX}.$$

Using the data of our problem, the equations become,

$$1.0000\beta_{XA} + .4146\beta_{XB} = .6747 \qquad (1)$$
$$.4146\beta_{XA} + 1.0000\beta_{XB} = .8029. \qquad (2)$$

These equations may be solved for β_{XA} and β_{XB} by any of the usual methods. For example, copy down the second equation, and multiply the first by .4146, thus;

$$.4146\beta_{XA} + 1.0000\beta_{XB} = .8029$$
$$.4146\beta_{XA} + .1719\beta_{XB} = .2797$$

Subtracting: $.8281\beta_{XB} = .5232$

Dividing by .8281: $\beta_{XB} = .6318$

Substituting in the first equation:

$$\beta_{XA} + .4146 \times .6318 = .6747$$

Solving: $$\beta_{XA} = .4126$$

The correctness of the solution may be tested by substitution in the second equation.

We now calculate the "multiple correlation coefficient", denoted by the symbol R, by means of the formula

$$R^2 = \beta_{XA} r_{AX} + \beta_{XB} r_{BX}.$$

Substituting: $R^2 = .4126 \times .6747 + .6318 \times .8029 = .7856$

Therefore, $R = .89$

Parenthetically, it may be observed, that while the subscripts of the r's may be interchanged, those of the β's may not. Thus β_{AX} does not denote the same number as β_{XA}. The meaning of β_{AX} will be explained later.

A second parenthetic observation will be of interest to many readers. These β's are the same as the "path coefficients" used by Sewall Wright in "Correlation and Causation" (Jour. Ag. Res. Vol. XX, No. 7, pp. 557-575), and the products, $\beta_{XA} r_{AX}$ and $\beta_{XB} r_{BX}$ are his "coefficients of determination".

Before discussing the meaning of R, we shall complete the calculations by computing the constants of the new regression equation from the formula:

$$\overline{X} = M_X + \beta_{XA} \frac{\sigma_X}{\sigma_A} (A - M_A) + \beta_{XB} \frac{\sigma_X}{\sigma_B} (B - M_B).$$

The use of the β's is now obvious. They play the same role in the multiple regression equation as the r's do in simple regression equations. Substituting our values for the above symbols, we find the new estimated value of X to be,

$$\overline{X} = 198.20 + .4126 \times \frac{61.23}{3.70} (A - 37.48) + .6318 \frac{61.23}{5.97} (B - 19.52)$$

or, $$\overline{X} = 6.828A + 6.478B - 184.16$$

As the final step in the calculations, we must now find the estimated values of X corresponding to each pair of actual values of A and B. For example, in Allamakee county, A = 40 bu. per acre corn yield, and B = 11% of farm land in small grain. Substituting in the regression equation, we find the corresponding estimated land value to be,

$$\overline{X} = 6.828 \times 40 + 6.478 \times 11 - 184.16 = \$160.22 \text{ per acre.}$$

Continuing this process for each county, we get Table 5.

Before discussing the reasons for the more glaring errors of estimate, we shall return to the subject of the meaning and use of the multiple correlation coefficient, R. In the first place, R is the simple correlation coefficient between actual land values and land values estimated from the regression equation. In other words, it is precisely the same kind of a measure of our success in estimating or predicting with two independent variables as r was a measure of our success with one independent variable (see page 15). In the second place, just as in simple correlation, R enables us to compute readily the standard error of estimate which, with two independent variables, A and B, is denoted by the symbol, $\sigma_{X \cdot AB}$.

The formula is $\sigma_{X \cdot AB} = \sigma_X \sqrt{1 - R^2}$, or in our example,

$$\sigma_{X \cdot AB} = 61.23 \sqrt{1 - .7856} = 61.23 \times .463 = \$28.35 \text{ per acre}$$

That is, our value R = .89 enables us to reduce the standard deviation of estimated values to 46.3% of the standard deviation of the X's; or in other words, to reduce it by 53.7%. (Compare

TABLE 5. LAND VALUE ESTIMATED FROM CORN YIELD AND FARM LAND IN SMALL GRAIN

Observation	Actual Average land value per acre	Estimated land value per acre	Error of Estimate
1. Allamakee	$ 87	$160	—73
2. Bremer	133	146	—13
3. Butler	174	171	3
4. Calhoun	285	310	—25
5. Carroll	263	244	19
6. Cherokee	274	252	22
7. Dallas	235	231	4
8. Davis	104	86	18
9. Fayette	141	146	— 5
10. Fremont	208	158	50
11. Howard	115	137	—22
12. Ida	271	238	33
13. Jefferson	163	159	4
14. Johnson	193	180	13
15. Kossuth	203	231	—28
16. Lyon	279	276	3
17. Madison	179	152	27
18. Marshall	244	246	— 2
19. Monona	165	178	—13
20. Pocahontas	257	283	—26
21. Polk	252	238	14
22. Story	280	239	41
23. Wapello	167	158	9
24. Warren	168	158	10
25. Winneshiek	115	178	—63

Table 4). We are now likely to find more than 95% (only 90% in this particular example) of our errors of estimate lying within the comparatively small range from $-\$56.70$ to $+\$56.70$ ($\pm 2\sigma_{X \cdot AB}$). In the third place, the standard deviation of R is (as in the case of r)

$$\sigma_R = \frac{1 - R^2}{\sqrt{n}} = \frac{1 - .7856}{5} = .04$$

which means that similarly derived R's are almost certain to lie above .80, thus practically always reducing the standard deviation of estimated values more than 40% (see Table 4).

Returning to a consideration of Table 5 we find that the new estimate of land value in Lyon county is very close to the actual value, whereas the estimate for Kossuth county is not so good as before. Kossuth, having about average corn yield and average land value, is estimated quite closely from corn yield alone; but since it is close to the highest county in percentage of land in small grain, the inclusion of the latter variable raises the land value estimate too much. Other factors will have to be introduced to counterbalance this effect. Land values in Allamakee and Winneshiek counties, especially in the case of Allamakee, are better than when only one independent variable was used. It is still necessary to take into account the fact that these counties have low percentages of their farm land in corn. Fremont county is still much above its estimated value; in fact, we have made a poorer estimate with two independent variables than with one. This is because of Fremont's unusually large percentage of farm land in corn, a characteristic which we shall certainly have to take into account before our problem is completed. In six other counties besides Fremont, our newly estimated values are not so close to actual values as was the case when corn yield alone was considered. In the other eighteen counties, our estimations are closer to the facts.

PART III.. MULTIPLE CORRELATION—MORE THAN THREE VARIABLES.

It is quite evident from what precedes that, while we have made progress in our attempt to analyze the relations between land value and associated variables, we are still far from a satisfactory knowledge of these relations. We shall complete our illustrative example by including three more variables, as follows: average number of improved acres per farm, C; number of brood sows per 1,000 acres, D; and percentage of farm land in corn, E. Table 6 gives the complete data for 25 Iowa counties.

The principles involved in handling more than three variables are identical with those explained in the three variable problem.

TABLE 6. DATA FROM 25 IOWA COUNTIES

Observation Number	County	Corn yield per acre 1910-1919	% farm land in small grain	No. improved acres per farm	No. brood sows per 1,000 acres	% farm land in corn	Value per acre of land Jan. 1, 1920	Sum.
		A	B	C	D	E	X	S
1	Allamakee	40	11	103	42	14	$ 87	297
2	Bremer	36	13	102	58	30	133	372
3	Butler	34	19	137	53	30	174	447
4	Calhoun	41	33	160	49	39	285	607
5	Carroll	39	25	157	74	33	263	591
6	Cherokee	42	23	166	85	34	274	624
7	Dallas	40	22	130	52	37	235	516
8	Davis	31	9	119	20	20	104	303
9	Fayette	36	13	106	53	27	141	376
10	Fremont	34	17	137	59	40	208	495
11	Howard	30	18	136	40	19	115	358
12	Ida	40	23	185	95	31	271	645
13	Jefferson	37	14	98	41	25	163	378
14	Johnson	41	13	122	80	28	193	477
15	Kossuth	38	24	173	52	31	203	521
16	Lyon	38	31	182	71	35	279	636
17	Madison	34	16	124	43	26	179	422
18	Marshall	45	19	138	60	34	244	540
19	Monona	34	20	148	52	30	165	449
20	Pocahontas	40	30	164	49	38	257	578
21	Polk	41	22	96	39	35	252	485
22	Story	42	21	132	54	41	280	570
23	Wapello	35	16	96	41	23	167	378
24	Warren	33	18	118	38	24	168	399
25	Winneshiek	36	18	113	61	21	115	364

First, calculate the r's, then the β's, and from these, R and the regression equation. However, with more than three variables, it is desirable to adopt some labor saving, systematizing and accuracy promoting devices, and these will now be explained.

The first of these devices is the introduction of an extra variable, S, whose values are shown in the last column of Table 6. Each number in this column is merely the sum (hence, S) of the corresponding numbers of the other columns; thus, for Allamakee county,

$$S = 40 + 11 + 103 + 42 + 14 + 87 = 297$$

The sums, S, are handled exactly like values of a seventh variable. The relatively small amount of extra labor involved in handling S furnishes a perfect check on the accuracy of the calculations, and obviates the necessity of repeating them. The details of the use of S will be given in the proper places below.

THE SIMPLE CORRELATION COEFFICIENTS

The second of the new devices is merely a form (Tables 7a and 7b) for systematizing the method of calculating the r's and the σ's. (See mimeograph bulletin by Bradford B. Smith, "The Use of Punched Card Tabulating Equipment in Multiple

TABLE 7a. FORMULAS FOR CORRELATION COEFFICIENTS

	A	B	C	D	E	X	S
Sums	ΣA	ΣB	ΣC	ΣD	ΣE	ΣX	ΣS
Means	M_A	M_B	M_C	M_D	M_E	M_X	M_S
A_1	ΣA^2	ΣAB	ΣAC	ΣAD	ΣAE	ΣAX	ΣAS
A_2	$(\Sigma A)M_A$	$(\Sigma A)M_B$	$(\Sigma A)M_C$	$(\Sigma A)M_D$	$(\Sigma A)M_E$	$(\Sigma A)M_X$	$(\Sigma A)M_S$
A_3	$\Sigma A^2-(\Sigma A)M_A$	$\Sigma AB-(\Sigma A)M_B$	$\Sigma AC-(\Sigma A)M_C$	$\Sigma AD-(\Sigma A)M_D$	$\Sigma AE-(\Sigma A)M_E$	$\Sigma AX-(\Sigma A)M_X$	Check Here
A_4	$\sqrt{\Sigma A^2-(\Sigma A)M_A}$	$\sqrt{\Sigma A^2-(\Sigma A)M_A}\sqrt{\Sigma B^2-(\Sigma B)M_B}$	$\sqrt{\Sigma A^2-(\Sigma A)M_A}\div\sqrt{\Sigma C^2-(\Sigma C)M_C}$	$\sqrt{\Sigma A^2-(\Sigma A)M_A}\div\sqrt{\Sigma D^2-(\Sigma D)M_D}$	$\sqrt{\Sigma A^2-(\Sigma A)M_A}\div\sqrt{\Sigma E^2-(\Sigma E)M_E}$	$\sqrt{\Sigma A^2-(\Sigma A)M_A}\div\sqrt{\Sigma X^2-(\Sigma X)M_X}$	
B_1		ΣB^2	ΣBC	ΣBD	ΣBE	ΣBX	ΣBS
B_2		$(\Sigma B)M_B$	$(\Sigma B)M_C$	$(\Sigma B)M_D$	$(\Sigma B)M_E$	$(\Sigma B)M_X$	$(\Sigma B)M_S$
B_3		$\Sigma B^2-(\Sigma B)M_B$	$\Sigma BC-(\Sigma B)M_C$	$\Sigma BD-(\Sigma B)M_D$	$\Sigma BE-(\Sigma B)M_E$	$\Sigma BX-(\Sigma B)M_X$	Check Here
B_4		$\sqrt{\Sigma B^2-(\Sigma B)M_B}$	$\sqrt{\Sigma B^2-(\Sigma B)M_B}\div\sqrt{\Sigma C^2-(\Sigma C)M_C}$	$\sqrt{\Sigma B^2-(\Sigma B)M_B}\div\sqrt{\Sigma D^2-(\Sigma D)M_D}$	$\sqrt{\Sigma B^2-(\Sigma B)M_B}\div\sqrt{\Sigma E^2-(\Sigma E)M_E}$	$\sqrt{\Sigma B^2-(\Sigma B)M_B}\div\sqrt{\Sigma X^2-(\Sigma X)M_X}$	
C_1			ΣC^2	ΣCD	ΣCE	ΣCX	ΣCS
C_2			$(\Sigma C)M_C$	$(\Sigma C)M_D$	$(\Sigma C)M_E$	$(\Sigma C)M_X$	$(\Sigma C)M_S$
C_3			$\Sigma C^2-(\Sigma C)M_C$	$\Sigma CD-(\Sigma C)M_D$	$\Sigma CE-(\Sigma C)M_E$	$\Sigma CX-(\Sigma C)M_X$	Check Here
C_4			$\sqrt{\Sigma C^2-(\Sigma C)M_C}$	$\sqrt{\Sigma C^2-(\Sigma C)M_C}\div\sqrt{\Sigma D^2-(\Sigma D)M_D}$	$\sqrt{\Sigma C^2-(\Sigma C)M_C}\div\sqrt{\Sigma E^2-(\Sigma E)M_E}$	$\sqrt{\Sigma C^2-(\Sigma C)M_C}\div\sqrt{\Sigma X^2-(\Sigma X)M_X}$	
D_1				ΣD^2	ΣDE	ΣDX	ΣDS
D_2				$(\Sigma D)M_D$	$(\Sigma D)M_E$	$(\Sigma D)M_X$	$(\Sigma D)M_S$
D_3				$\Sigma D^2-(\Sigma D)M_D$	$\Sigma DE-(\Sigma D)M_E$	$\Sigma DX-(\Sigma D)M_X$	Check Here
D_4				$\sqrt{\Sigma D^2-(\Sigma D)M_D}$	$\sqrt{\Sigma D^2-(\Sigma D)M_D}\div\sqrt{\Sigma E^2-(\Sigma E)M_E}$	$\sqrt{\Sigma D^2-(\Sigma D)M_D}\div\sqrt{\Sigma X^2-(\Sigma X)M_X}$	
E_1					ΣE^2	ΣEX	ΣES
E_2					$(\Sigma E)M_E$	$(\Sigma E)M_X$	$(\Sigma E)M_S$
E_3					$\Sigma E^2-(\Sigma E)M_E$	$\Sigma EX-(\Sigma E)M_X$	Check Here
E_4					$\sqrt{\Sigma E^2-(\Sigma E)M_E}$	$\sqrt{\Sigma E^2-(\Sigma E)M_E}\sqrt{\Sigma X^2-(\Sigma X)M_X}$	
X_1						ΣX^2	ΣXS
X_2						$(\Sigma X)M_X$	$(\Sigma X)M_S$
X_3						$\Sigma X^2-(\Sigma X)M_X$	Check Here
X_4						$\sqrt{\Sigma X^2-(\Sigma X)M_X}$	
Std Dev.	σ_A	σ_B	σ_C	σ_D	σ_E	σ_X	

TABLE 7b. CALCULATION OF CORRELATION COEFFICIENTS

	A	B	C	D	E	X	S
Sums	937	488	3342	1361	745	4955	11820
Means	37.48	19.52	133.68	54.44	29.80	198.20	473.12
A_1	35461	18519	125886	51762	28261	189533	449422
A_2	35119	18290	125258	51010	27923	185713	443313
A_3	342	229	628	752	338	3820	6109
A_4	18.493	552.31	2475.8	1505.6	641.7	5661.8	
B_1		10418	68242	27397	15242	104064	243882
B_2		9526	65236	26567	14542	96722	230883
B_3		892	3006	830	700	7342	12999
B_4		29.866	3990.5	2431.5	1036.3	9143.8	
C_1			464684	188152	101900	688739	1637603
C_2			446759	101930	99592	662384	1581167
C_3			17925	6214	2308	26355	56436
C_4			133.88	10900	4645.5	40989	
D_1				80721	41638	282889	672559
D_2				74093	40558	269750	643916
D_3				6628	1080	13139	28643
D_4				81.413	2824.9	24925	
E_1					23405	156817	367263
E_2					22201	147659	352474
E_3					1204	9158	14789
E_4					34.699	10623	
X_1						1075817	2497859
X_2						982081	2344310
X_3						93736	153549
X_4						306.16	
Std Dev	3.70	5.97	26.78	16.28	6.94	61.23	

Correlation Problems", Bureau of Agricultural Economics, Washington, D. C., October, 1923.) In Table 7a the entries are indicated by symbols alone, just as in a formula. In Table 7b are shown the corresponding numbers in our problem. By comparing these tables with the formulas and numbers previously used the familiar parts will be quickly located, and the whole scheme easily comprehended. The blank spaces in the lower left portion of the tables are caused by the elimination of unnecessary repetition. For example, line B_1 column A would naturally contain ΣBA, but this is identical with ΣAB (18,519) and is therefore omitted.

In Table 7b, the sums in the first line are recorded directly from the calculating machine, all original data being found in Table 6. The correctness of these sums is checked in this line by observing that the sum of the first six of them is equal to the seventh; that is,

$$937 + 488 + 3{,}342 + 1{,}361 + 745 + 4{,}955 = 11{,}828$$

The "product moments" (including the sums of the squares) in lines A_1, B_1, etc., are recorded directly from the machine. The calculation may be facilitated by folding Table 6 vertically so as to bring into juxtaposition the pair of numbers to be multiplied.

The check in line A_1 is furnished by adding the first six numbers in that line. The sum should be the same as the product moment (ΣAS) already recorded under S in the same line; that is,

$$35{,}461 + 18{,}519 + \text{etc.} = 449{,}422$$

To check line B_1 it must be remembered that the first of the product moments ($\Sigma BA = \Sigma AB = 18{,}519$) is omitted. It is necessary, therefore, to start at the top of column B, come down to line B_1, then go across the line; thus,

$$18{,}519 + 10{,}418 + 68{,}242 + \text{etc.} = 243{,}882$$

Similarly, start at the top of column C, go down to line C_1, then across, obtaining

$$125{,}886 + 68{,}242 + 464{,}684 + 188{,}152 + 101{,}900 + 688{,}739 = 1{,}637{,}603$$

The products in lines A_2, B_2, etc., are also recorded directly from the machine, the data being found in the first and second lines of the present table. The check is the same as before, except that all the numbers checked are now found in lines with subscripts 2. For line A_2 we have,

$$35{,}119 + 18{,}290 + 125{,}258 + \text{etc.} = 443{,}313$$

For line E_2 (and column E),

$$27{,}923 + 14{,}542 + 99{,}592 + 40{,}558 + 22{,}201 + 147{,}659 = 352{,}474$$

Each number in line A_2 is subtracted from the number just above it in line A_1, the results appearing in line A_3. The same relations obtain in lower parts of the tables. It happens in this problem that all these differences are positive; but in another problem the lower number might in some places be larger, in which case the difference would be negative. This would result in a negative correlation coefficient. The check in lines with subscripts 3 is the same as before, and is the final check on this part of the calculation. With skillful calculators, either or both the preceding checks may be omitted, but this last one is essential.

The number of significant figures to which the results check depends, of course, upon the number of figures carried in the means. In the illustrative problem, since the number of observations is 25, the means are made arithmetically exact by carrying only two decimal places. In another problem, however, if it is desired to check results to seven significant figures (as is done in this problem) seven figures would have to be carried in the means and even then the last figures would not usually check, as is the case in the last two numbers of the illustrative problem. Of course, the extra figures have no statistical significance, and their use would in any case be merely a matter of office practice. So far as statistical significance is concerned, all the numbers used in this problem might have been limited to the first three, or at most four figures.

The first number in line A_4 (18.493) is the square root of the number just above it (342); and similarly, for the first numbers in lines B_4, etc. Each of these square roots when divided by the square root of the number of observations (in our problem $\sqrt{25}$) gives the corresponding standard deviation σ in the bottom row of the table.

The remaining numbers in lines A_4, B_4, etc., are products of two square roots, namely the first square root in the same line by the last square root in the same column. For example, the number (1,036.3) in line B_4 column E is the product of 29.866 (line B_4 column B) by 34.699 (line E_4 column E).

The correlation coefficients of A with each of the remaining variables (not including S) are calculated by dividing each number in line A_3 (not including column A) by the number just below it. As an example, from column E,

$$r_{AE} = \frac{338}{641.7} = .5267$$

As an example of the similar use of later rows, consider rows C_3 and C_4, column X; from these we obtain,

$$r_{CX} = \frac{26{,}355}{40{,}989} = .6430$$

THE NORMAL EQUATIONS

In order to record the r's and use them for calculating the β's in the simplest way, we now turn to a consideration of Tables 8a and 8b. Here, as its value is calculated, each r is recorded in the row and column corresponding to its subscripts. The exact position of each r is clearly indicated in Table 8a. Observe that $r_{AA} = 1.0000$, $r_{BB} = 1.0000$, etc. These tables exhibit the third and last of the new devices to be considered in this part. This device is a short scheme for obtaining the solution of the normal equations, a set of simultaneous, linear equations having the same number of "unknowns" (five in our illustrative problem) as there are independent variables. The unknowns in these normal equations are the partial regression coefficients, β_{XA}, β_{XB}, β_{XC}, etc.

Written out in full, these five normal equations appear thus:

$$\begin{aligned}
\beta_{XA} + r_{AB}\beta_{XB} + r_{AC}\beta_{XC} + r_{AD}\beta_{XD} + r_{AE}\beta_{XE} &= r_{AX} \\
r_{BA}\beta_{XA} + \beta_{XB} + r_{BC}\beta_{XC} + r_{BD}\beta_{XD} + r_{BE}\beta_{XE} &= r_{BX} \\
r_{CA}\beta_{XA} + r_{CB}\beta_{XB} + \beta_{XC} + r_{CD}\beta_{XD} + r_{CE}\beta_{XE} &= r_{CX} \\
r_{DA}\beta_{XA} + r_{DB}\beta_{XB} + r_{DC}\beta_{XC} + \beta_{XD} + r_{DE}\beta_{XE} &= r_{DX} \\
r_{EA}\beta_{XA} + r_{EB}\beta_{XB} + r_{EC}\beta_{XC} + r_{ED}\beta_{XD} + \beta_{XE} &= r_{EX}
\end{aligned}$$

(See Kelley: "Statistical Method", p. 296.)

It will be observed that there is a diagonal row of β's through this array of equations, from the upper left to the lower right corner, each of whose coefficients is unity. If, now, we remember that $r_{AB} = r_{BA}$, $r_{AC} = r_{CA}$, $r_{CE} = r_{EC}$, etc., it can be seen that the r's in the upper right hand part of the array and the equal r's in the lower left hand part are arranged symmetrically with respect to the diagonal of unity coefficients. It is for this reason that short methods of solution can be used, and that we need keep only that portion of the equations above and to the right of the diagonal, together with the "1's" in the diagonal itself. Finally, it is unnecessary to record the β's, since only the r's are required for calculation. Thus we get the arrangement of the r's in tables 8a and 8b. The directions to be given for manipulations in these tables have as their objective the solution of the normal equations, giving finally the values of the β's. For an

extensive explanation of the whole process, see Wright and Hayford: "Adjustment of Observations", pp. 114-120.

Most of the details of manipulation can be understood by study of the directions given in the tables themselves, and comparison of the two tables. Each symbol in Table 8a stands for whatever number might be entered in the corresponding cell in any particular problem. Thus, in our problem, r_{AC} (line 1, column C) stands for the number, .2536; [bb] (line 5, column B) stands for .8281; [dx] (line 17, column X) stands for — .1199; and — [dx] (line 2 of reverse, column X) stands for .1199. Some statements of general principles will help the operator to carry the details in mind.

First. Each block of lines is narrower by one column than the preceding block, and after the B-block each block of lines is one line wider than the preceding block. If there were an F variable, the F-block would contain 8 lines (25 to 32 inclusive) and so on for any number of variables.

Second. Beginning with the B-block, the next to the last line in each block (lines 5, 10, 16, etc.) consists of the algebraic sums of all the entries above it in the same block. Thus in Table 8b the number (.8281) in line 5, column B, is equal to 1 — .1719, and the number (.0402) in line 16, column E, is equal to

$$.3824 - .2631 - .0741 - .0050$$

Third. The sums in the next-to-the-last line of each block (lines 5, 10, 16, 23) are each to be divided by the first such sum in the same block, the signs reversed, and the quotients entered just below the dividends. Thus, in line 10 the divisor is .4307; the dividends are .4307, .3385, .0063, .0633 and .8388; and the quotients with signs changed appearing in line 11 are — 1.0000, — .7859, — .0146, — .1470 and — 1.9475.

Fourth. Each of the remaining lines in any block consists of products calculated from one of the preceding blocks. Thus, in block D the products in line 13 are calculated from the A-block, those in line 14 from the B-block and those in line 15 from the C-block. To illustrate from Table 8b, consider the products in line 14, block D. These products come from the B-block, thus:

	Line	D	E	X	S
Multiplicands	5	.1343	.4571	.5232	2.5894
Multiplier	6	—.1622			
Products	14	—.0218	—.0741	—.0849	— .4200

TABLE 8a. DIRECTIONS FOR SOLVING NORMAL EQUATIONS

DESCRIPTION	Block	Line	A	B	C	D	E	X	S
Enter A correlation coefficients	A	1	1	r_{AB}	r_{AC}	r_{AD}	r_{AE}	r_{AX}	↔ sum
Change Signs		2	-1	$-r_{AB}$	$-r_{AC}$	$-r_{AD}$	$-r_{AE}$	$-r_{AX}$	
Enter B correlation coefficients	B	3		1	r_{BC}	r_{BD}	r_{BE}	r_{BX}	↳ sum
Multiply Line 1 by [r_{AB}] Line 2, Col B		4		[r_{AB}] x r_{AB}	[r_{AB}] x r_{AC}	[r_{AB}] x r_{AD}	[r_{AB}] x r_{AE}	[r_{AB}] x r_{AX}	[r_{AB}] x [sum]
Add Lines 3 and 4		5		[bb]	[bc]	[bd]	[be]	[bx]	[bs]
Divide Line 5 by [bb], and change signs		6		-1	[bc]	[bd]	[be]	[bx]	Check here
Enter C correlation coefficients	C	7			1	r_{CD}	r_{CE}	r_{CX}	↳ sum
Multiply Line 1 by [r_{AC}], Line 2, Col C.		8			[r_{AC}] x r_{AC}	[r_{AC}] x r_{AD}	[r_{AC}] x [r_{AE}	[r_{AC}] x r_{AX}	[r_{AC}] x [sum]
" " 5. [bc], Line 6, " C.		9			[bc] x [bc]	[bc] x [bd]	[bc] x [be]	[bc] x [bx]	[bc] x [bs]
Add Lines 7, 8 and 9.		10			[cc]	[cd]	[ce]	[cx]	[cs]
Divide Line 10 by [cc] and change signs		11			-1	[cd]	[ce]	[cx]	Check here
Enter D correlation coefficients	D	12				1	r_{DE}	r_{DX}	↳ sum
Multiply Line 1 by [r_{AD}], Line 2, Col. D		13				[r_{AD}] x r_{AD}	[r_{AD}] x r_{AE}	[r_{AD}] x r_{AX}	[r_{AD}] x [sum]
" " 5 - [bd], " 6, " D		14				[bd] x [bd]	[bd] x [be]	[bd] x [bx]	[bd] x [bs]
" " 10 - [cd], " 11, " D		15				[cd] x [cd]	[cd] x [ce]	[cd] x [cx]	[cd] x [cs]
Add Lines 12-15		16				[dd]	[de]	[dx]	[ds]
Divide Line 16 by [dd] and change signs		17				-1	[de]	[dx]	Check here
Enter E correlation coefficients	E	18					1	r_{EX}	↳ sum
Multiply Line 1 by [r_{AE}], Line 2, Col E.		19					[r_{AE}] x r_{AE}	[r_{AE}] x r_{AX}	[r_{AE}] x [sum]
" " 5 . [be], " 6 . E		20					[be] x [be]	[be] x [bx]	[be] x [bs]
" " 10 . [ce], " 11 . E		21					[ce] x [ce]	[ce] x [cx]	[ce] x [cs]
" " 16 . [de], " 17 . E		22					[de] x [de]	[de] x [dx]	[de] x [ds]
Add Lines 18-22		23					[ee]	[ex]	[es]
Divide Line 23 by [ee] and change signs		24					-1	[ex]	Check here
$\beta_{xE} = -[ex]$	Reverse	1					β_{xE}	$-[ex]$	
β_{xD} = sum of 2 terms at its right.		2				β_{xD}	β_{xE} x [de]	$-[dx]$	
β_{xC} . . . 3		3			β_{xC}	β_{xD} x [cd]	β_{xE} x [ce]	$-[cx]$	
β_{xB} . . . 4		4		β_{xB}	β_{xC} x [bc]	β_{xD} x [bd]	β_{xE} x [be]	$-[bx]$	
β_{xA} . . . 5		5	β_{xA}	β_{xB} x ($-r_{AB}$)	β_{xC} x ($-r_{AC}$)	β_{xD} x ($-r_{AD}$)	β_{xE} x ($-r_{AE}$)	r_{AX}	

TABLE 8b. SOLUTION OF NORMAL EQUATIONS

Description	Block	Line	A	B	C	D	E	X	5
Enter A correlation coefficients	A	1	1.0000	.4146	.2536	.4995	.5267	.6747	3.3691
Change signs		2		-.4146	-.2536	-.4995	-.5267	-.6747	
Enter B correlation coefficients	B	3		1.0000	.7518	.3414	.6755	.8029	3.9862
Multiply Line 1 by (-.4146), Line 2, Col. B		4		-.1719	-.1051	-.2071	-.2184	-.2797	-1.3968
Add Lines 3 and 4		5		.8281	.6467	1.1343	.4571	.5232	2.5894
Divide Line 5 by (.8281) and change signs		6		-1.0000	-.7809	-1.3622	-.5520	-.6318	-3.1269
Enter C correlation coefficients	C	7			1.0000	.5701	.4968	.6430	3.7153
Multiply Line 1 by (-.2536), Line 2, Col C		8			-.0643	-1.1267	-.1336	-.1711	-.8549
" " 5 (-.7809), " 6 " C		9			-.5050	-1.1049	-.3569	-.4086	-2.0221
Add Lines 7, 8 and 9		10			.4307	.3385	.0063	.0633	.8388
Divide Line 10 by (.4307) and change signs		11			-1.0000	-.7859	-.0146	-.1470	-1.9475
Enter D correlation coefficients	D	12				1.0000	.3824	.5271	3.3205
Multiply Line 1 by (-.4995) Line 2, Col. D		13				-.2495	-.2631	-.3370	-1.6829
" " 5 (-.1622), " 6 " D		14				-.0218	-.0741	-.0849	-.4200
" " 10 (-.7859), " 11 " D		15				-.2660	-.0050	-.0497	-.6592
Add Lines 12-15		16				.4627	.0402	.0555	.5584
Divide Line 16 by (.4627) and change signs		17				-1.0000	-.0869	-.1199	-1.2068
Enter E correlation coefficients	E	18					1.0000	.8621	3.9435
Multiply Line 1 by (-.5267) Line 2, Col E		19					-.2774	-.3554	-1.7745
" " 5 (-.5520), " 6, " E		20					-.2523	-.2888	-1.4293
" " 10 (-.0146), " 11, " E		21					-.0001	-.0009	-.0122
" " 16 (-.0869), " 17, " E		22					-.0035	-.0048	-.0485
Add Lines 18-22		23					.4667	.2122	.6790
Divide Line 23 by (.4667) and change signs		24					-1.0000	-.4547	-1.4549
β_{xe} = .4547	Reverse	1					.4547	.4547	
β_{xd} = -.0395 +.1199 -.0804		2				.0804	-.0395	.1199	
β_{xc} = -.0632 -.0066 +.1470 -.0772		3			.0772	-.0632	-.0066	.1470	
β_{xb} = -.0603 -.0130 -.2510 +.6318 = .3075		4		.3075	-.0603	-.0130	-.2510	.6318	
β_{xa} = -.1275 -.0196 -.0402 -.2395 +.6747 = .2479		5	.2479	-.1275	-.0196	-.0402	-.2395	.6747	

Fifth. The last line in each block contains the coefficients with signs reversed of an equation from which some of the unknown β's have been eliminated. Thus, from line 17 we may infer that

$$1.0000\beta_{XD} + .0869\beta_{XE} = .1199$$

Similarly from line 24,

$$\beta_{XE} = .4547$$

which is, therefore, the first one of the β's whose value is found after all the rest of them have been eliminated from the equations.

Sixth. The S-column furnishes a check on the accuracy of the work in each block, but does not check the calculations of the r's. The entries in the S-column are not carried over from Table 7. That in line 1 is simply the sum of the r's to the left of it in the same line; that is,

$$1.0000 + .4146 + .2536 + .4995 + .5267 + .6747 = 3.3691$$

The entry in line 3, column S is likewise the sum of five r's arranged in the same "down and across" manner as used in Table 7b; thus, going down column B and across line 3, we have

$$.4146 + 1.0000 + .7518 + .3414 + .6755 + .8029 = 3.9862$$

As a final illustration, consider the entry in line 12, column S. Down column D and across line 12,

$$.4995 + .3414 + .5701 + 1.0000 + .3824 + .5271 = 3.3205$$

After the entries are made in column S, they are treated exactly like the original entries in the other columns. (See Table 8a, column S.) The check is furnished in the last line of each block. The number in the S column of that line should be (approximately) equal to the sum of the numbers to the left of it in the same line (*not down and across*). Consider, for example, line 11,

$$-1.0000 - .7859 - .0146 - .1470 = -1.9475$$

Seventh. The "Reverse" (bottom five lines of tables 8a and 8b) is the process of finding the values of the preceding β's by retracing our steps, equation by equation. Some of the details will now have to be explained, as follows:

(1) In column X, copy in reverse order with sign changed the last number (in the same column) in each block above. This is clearly indicated in Table 8a, column X of the reverse.

(2) In line 1, column E, copy the value of β_{XE}, which in our problem is .4547. The two numbers in reverse, line 1, are always the same.

(3) In column E, below β_{XE}, enter in reverse order the prod-

ucts of β_{XE} by the last number appearing in that column in each of the blocks above the E-block. For example, $-.0395 = .4547 \times -.0869$, and $-.0066 = .4547 \times -.0146$.

(4) In reverse line 2, add (algebraically) the numbers in columns X and E (.1199 — .0395), placing the sum (.0804) in the same line, column D. This sum is the value of β_{XD}. It may easily be seen that the operations in reverse line 2 result in the substitution of the value of β_{XE} in an equation mentioned above, namely

$$1.0000\beta_{XD} + .0869\beta_{XE} = .1199$$

and also in its solution for the value of β_{XD};

$$\beta_{XD} = .1199 - (.0869 \times .4547) = .1199 - .0395 = .0804$$

(5) Repeat in column D the operations just described, using as multiplier the value of β_{XD} (.0804). We now compute in reverse line 3,

$$\beta_{XC} = .1470 - .0066 - .0632 = .0772$$

What we have really done in reverse line 3 is to substitute the values of β_{XE} and β_{XD} in an equation inferred from line 11 above, as follows:

$$1.0000\beta_{XC} + .7859\beta_{XD} + .0146\beta_{XE} = .1470$$

and solve the resulting equation for β_{XC}.

Continue this reverse process until all the β's have been calculated, then verify the results by substituting their values in some one of the original normal equations. For example, reading down column D to line 12, then along line 12, we infer the equation,

$$.4995\beta_{XA} + .3414\beta_{XB} + .5701\beta_{XC} + 1.0000\beta_{XD} + .3824\beta_{XE} = .5271$$

Substituting the values of the β's in the left member of this equation it becomes,

$$(.4995 \times .2479) + (.3414 \times .3075) + (.5701 \times .0772) + (.0804) + (.3824 \times .4547)$$

Without clearing the machine, we compute the sum of these products as .5271, thus verifying the correctness of the β's in the equation above. For verification purposes, any of the original normal equations may be used *except the first (line 1)*, which has already been made use of (reverse, line 5) for calculating the value of β_{XA}.

THE MULTIPLE CORRELATION COEFFICIENT

We are now ready to calculate the multiple correlation coefficient, R, from the equation

$$R^2 = \beta_{XA} \cdot r_{AX} + \beta_{XB} \cdot r_{BX} + \text{etc.},$$
$$= (.2479 \times .6747) + (.3075 \times .8029) + (.0772 \times .6430) + (.0804 \times .5271) + (.4547 \times .8621)$$

The machine gives directly the sum of these products as .8982. The factors are readily found in Table 8b; and after a little practice the multiplications and additions may be carried through on the machine without making such a list as that given in the equation above, without clearing the machine. Finally, $R = \sqrt{.8982} = .95$

THE STANDARD ERROR OF ESTIMATE

This value of R shows that if we attempt to estimate land values from these five independent variables, the standard error of estimate will be

$$\sigma_{X \cdot ABCDE} = \sigma_X \sqrt{1 - R^2} = .319\sigma_X \text{ or } 31.9\% \text{ of } \sigma_X.$$

That is,

$$\sigma_{X \cdot ABCDE} = .319 \times \$61.23 = \$19.53$$

Thus, we have reduced the original standard deviation by 68.1%. In Part II, we found that by using two independent variables we could reduce the original standard deviation by only 53.7%. The addition of three more independent variables is therefore of real value. On the other hand, the fact that the standard error of estimate is still 31.9% of σ_X shows that the problem is not completely solved. There are other influences on the price of land which have not been considered, and it is the search for these that will engage the interest of the student of economics.

THE REGRESSION EQUATION

The regression equation with five independent variables is

$$\bar{X} = M_X + \beta_{XA} \cdot \frac{\sigma_X}{\sigma_A} (A - M_A) + \beta_{XB} \cdot \frac{\sigma_X}{\sigma_B} (B - M_B) + \text{etc.}$$

With our data, this becomes

$$\bar{X} = 198.20 + .2479 \times \frac{61.23}{3.70} (A - 37.48) + .3075 \times \frac{61.23}{5.97} (B - 19.52) + .0772 \times \frac{61.23}{26.78} (C - 133.68) + .0804 \times \frac{61.23}{16.28} (D - 54.44) + .4547 \times \frac{61.23}{6.94} (E - 29.80)$$

TABLE 9. LAND VALUE ESTIMATED FROM FIVE VARIABLES

County	Average land value per acre	Estimated land value per acre	Error of Estimate
Allamakee	$ 87	$109	—22
Bremer	133	168	—35
Butler	174	183	— 9
Calhoun	285	295	—10
Carroll	263	245	18
Cherokee	274	260	14
Dallas	235	244	— 9
Davis	104	86	18
Fayette	141	155	—14
Fremont	208	219	—11
Howard	115	115	..
Ida	271	246	25
Jefferson	163	149	14
Johnson	193	191	2
Kossuth	203	225	—22
Lyon	279	271	8
Madison	179	152	27
Marshall	244	247	— 3
Monona	165	188	—23
Pocahontas	257	278	—21
Polk	252	230	22
Story	280	266	14
Wapello	167	139	28
Warren	168	144	24
Winneshiek	115	150	—35

$$\overline{X} = 4.103A + 3.154B + .1766C + .3022D + 4.012E - 176.76$$

Using this equation, we calculate the land values shown in Table 9.

It should be observed that in this problem where all the β's are positive, the land value of any one county is found by adding five products and subtracting $176.76. This should be done, as usual, without clearing the machine. In this way, the estimated values for all the counties can be found in a short time. If part of the β's were negative, their terms should be subtracted instead of added. This is done in the crank driven machine by turning the crank backward (subtracting) instead of forward. In the key driven machine, the "complementary" number must be used.

If we compare our latest errors of estimate with those made on the basis of two independent variables, we find notable improvement in the cases of Allamakee, Calhoun, Fremont, Howard, Story and Winneshiek counties, but much poorer estimates for Bremer and Wapello. Land values in nine more of the twenty-five counties are not estimated so well with these five independent variables as with two, but the changes are rela-

tively insignificant. This strengthens our previous conclusion that the student of economics has still a long way to go before he finds all the factors that are highly associated with land values.

SCORING

The regression equation is the best scoring device available, the "score" of any individual being the value of the criterion, X, calculated from a given set of values of the independent variables. In this sense, the estimated land values found in the second column of Table 9 constitute the scores of the corresponding counties on the basis of land value.

There are times, however, when a simpler scoring device is desirable. While there is no general agreement on the subject, the partial regression coefficients probably constitute the simplest and most straightforward data for making a score card. In the table below is entered the value of each β in our land value problem, and just beneath it is placed a rate percent. Each rate percent is found by dividing the corresponding β by the sum (1.1677) of the five β's.

SCORE CARD

	Corn Yield	Small Grain	Improved Land	Brood Sows	Corn Land	Sum.
Coefficients	.2479	.3075	.0772	.0804	.4547	1.1677
Rate Percents or Scores	21	26	7	7	39	100

If the counties of Iowa are to be scored on the basis of the data in Table 6, it thus appears that 39% of the score should be based on the percentage of farm land in corn, 26% on percentage of farm land in small grain, 21% on corn yield in bushels per acre, and 7% each on number of acres of improved land per farm and number of brood sows per 1,000 acres.

PART IV. PARTIAL CORRELATION COEFFICIENTS

In partial correlation coefficients, the attempt is made to determine the degree of association that would exist between two variables if we could eliminate the effects of their common associations with other variables. For example, consider the correlation coefficient .38 between the number of brood sows per 1,000 acres (D) and the percentage of farm land in corn (E) in the 25 counties which have been used as an illustration.

This is a statistically significant positive correlation, as shown by the fact that .38 is 2.23 times its own standard deviation. (We may determine from a table of the probability integral, such as Pearson's "Tables for Statisticians and Biometricians", Table II; or Pearl and Miner's table published as Table No. 40 in Pearl's "Medical Biometry and Statistics", the likelihood that nearly 99% of the correlation ratios calculated from similarly selected data would be greater than zero.)

The question arises, is $r_{DE} = .38$ because of some underlying relation between these variables, or merely because each of them is intimately associated with some other variable, such as price per acre of land (X)? We seek an answer in the "partial correlation coefficient between E and D independent of X", which we shall denote by the symbol $r_{DE \cdot X}$. The formula is

$$r_{DE \cdot X} = \frac{r_{DE} - r_{DX} \times r_{EX}}{\sqrt{1 - r_{DX}^2}\ \sqrt{1 - r_{EX}^2}}$$

In our example

$$r_{DE \cdot X} = \frac{.38 - .53 \times .86}{\sqrt{1 - (.53)^2}\ \sqrt{1 - (.86)^2}} = -.19$$

This means that if we could eliminate the common association of variables D and E with X, there would actually remain a small negative correlation between D and E; that is, independent of their common association with land value, there is a very slight tendency for large numbers of brood sows per 1,000 acres to be associated with small percentages of farm land in corn and vice versa.

In order to make clear the meaning of the partial correlation coefficient, we shall give two explanations as follows:

First. Imagine the number of counties in our problem increased to some large number such as a thousand, with no change in the simple and partial correlations discussed above. Consider a group of counties whose land values all lie in some such small interval as from $250 to $260 per acre. There might be 25 or more counties in such a group, and for practical purposes we could consider their land values to be the same. We could then calculate r_{DE} for this group, thus determining the degree of association between the number of brood sows per 1,000 acres and percentage of land in corn in counties having a common land value. This could be done for each other small group having a common land value. Then it may be said that the partial correlation coefficient $r_{DE \cdot X}$ would be a kind of average of all the simple correlation coefficients so obtained. For an illustra-

tion of this, see Pearl's "Medical Biometry and Statistics", pp. 322-25.

Second. In the three variable problem considered above, involving D, E and X, let us consider the estimation first of D from X, then of E from X. The two regression equations are written

$$\overline{D} = M_D + r_{DX}\frac{\sigma_D}{\sigma_X}(X - M_X),$$

and

$$\overline{E} = M_E + r_{EX}\frac{\sigma_E}{\sigma_X}(X - M_X),$$

as explained in Part I. After the values of $\overline{D}$ and $\overline{E}$ are calculated for each value of X, two groups of errors of estimate can be computed, $(D - \overline{D})$ and $(E - \overline{E})$. If these errors of estimate are arranged in pairs, one pair for each value of X, then it may be proved that the partial correlation coefficient $r_{DE \cdot X}$ is equal to the simple correlation coefficient between these pairs of errors of estimate. (See Kelley's "Statistical Method", pages 284-287.)

The explanations given above may be extended to partial correlation coefficients of higher orders. Thus, if we first calculate as above

$$r_{DE \cdot X} = -.19, \quad r_{AD \cdot X} = .22, \quad r_{AE \cdot X} = -.13$$

we may then find the partial correlation coefficient between D and E independent of both corn yield per acre (A) and land value (X) by means of the formula

$$r_{DE \cdot AX} = \frac{r_{DE \cdot X} - (r_{AD \cdot X})\ (r_{AE \cdot X})}{\sqrt{1 - r_{AD \cdot X}^2}\ \sqrt{1 - r_{AE \cdot X}^2}}$$

$$= \frac{-.19 - (.22)\ (-.13)}{\sqrt{1 - (.22)^2}\ \sqrt{1 - (-.13)^2}}$$

$$= -.17$$

According to the first explanation given above, this means that for groups having corn yields per acre the same, as well as land values the same, the average simple correlation between brood sows per 1,000 acres and percentage of land in corn would be negative ($-.17$), but not highly significant statistically. According to the second explanation, $-.17$ is the simple correlation coefficient between two series of errors of estimate: the first being the errors of estimate, $(D - \overline{D})$, made when we estimate

values of D from given values of A and X (Part II); and the second $(E - \overline{E})$, made when we estimate values of E from the same given values of A and X.

It is obvious from the foregoing that the calculations involved in multiple correlation though extensive are simple in form. A table giving values of $(1 - r^2)$ or $\sqrt{1 - r^2}$ for various values of r is highly desirable. John Rice Miner has calculated such tables. Publishers, John Hopkins Press. The calculations are quickly performed either with a machine or by means of a slide rule.

In most cases, however, a relatively brief extension of the calculations described in Part III will yield all the partial regression coefficients that are desired; namely, those of highest order giving the correlations between the criterion and the several independent variables. For example, in our six variable problem, we may wish the partial correlation coefficient between percentage of land in corn (E) and land value (X) independent of the other four variables. The symbol is $r_{EX \cdot ABCD}$. Its value can, of course, be obtained by building up the partials of lower orders according to the formulas already given, but the quicker method will now be explained.

If in Table 8b we should interchange the two columns E and X, as is done in Table 10, and make the necessary re-calculations in the last block (which is now the X-block), the last number in column E (line 24) with sign changed (1.0706) is easily seen to be the partial regression coefficient, β_{EX}. This is the coefficient that would be used in the regression equation if we were considering E as the dependent variable, and estimating E from X and the remaining four variables. β_{XE} (calculated in Part III) and β_{EX} are called "conjugate regression coefficients". We may now use the formula

$$\begin{aligned} r_{EX \cdot ABCD} &= \sqrt{\beta_{XE} \times \beta_{EX}} \\ &= \sqrt{.4547 \times 1.0706} \\ &= .6977 \end{aligned}$$

It has already been explained that the notation used for the β's in a six variable problem is quite inadequate. It should be observed that the complete notation for the above equation is

$$r_{EX \cdot ABCD} = \sqrt{\beta_{XE \cdot ABCD} \times \beta_{EX \cdot ABCD}}.$$

If we wish to calculate $r_{DX \cdot ABCE}$ we must have β_{DX} in addition to the β_{XD} previously calculated. To get β_{DX} we may interchange columns D and E in Table 10 with the corresponding change in block letters, and make the necessary re-calculations as in Table 11. The last number now appearing in column D (line 24) with sign changed (.3511) is β_{DX}. Then

$$r_{DX \cdot ABCE} = \sqrt{.0804 \times .3511} = .1679$$

TABLE 10

BLOCK	LINE	D	X	E	S
A	1	.4995	.6747	.5267	3.3691
	2	-.4995	-.6747	-.5267	
B	3	.3414	.8029	.6755	3.9862
	4	-.2071	-.2797	-.2184	-1.3968
	5	.1343	.5232	.4571	2.5894
	6	-.1622	-.6318	-.5520	-3.1269
C	7	.5701	.6430	.4968	3.7153
	8	-.1267	-.1711	-.1336	-.8544
	9	-.1049	-.4086	-.3569	-2.0221
	10	.3385	.0633	.0063	.8388
	11	-.1859	-.1470	-.0146	-1.9475
D	12	1.0000	.5271	.3824	3.3205
	13	-.2495	-.3370	-.2631	-1.6829
	14	-.0218	-.0849	-.0741	-.4200
	15	-.2660	-.0497	-.0050	-.6592
	16	.4627	.0555	.0402	.5584
	17	-1.0000	-.1199	-.0869	-1.2068
X	18		1.0000	.8621	4.5098
	19		-.4552	-.3554	-2.2731
	20		-.3306	-.2888	-1.6360
	21		-.0093	-.0009	-.1233
	22		-.0067	-.0048	-.0670
	23		.1982	.2122	.4104
	24		-1.0000	-1.0706	-2.0706
Reverse	1		1.0706	1.0706	
	2	-.0415	-.1284	.0869	

TABLE 11

BLOCK	LINE	E	X	D	S
A	1	.5267	.6747	.4995	3.3691
	2	-.5267	-.6747	-.4995	
B	3	.6755	.8029	.3414	3.9862
	4	-.2184	-.2797	-.2071	-1.3968
	5	.4571	.5232	.1343	2.5894
	6	-.5520	-.6318	-.1622	-3.1269
C	7	.4968	.6430	.5701	3.7153
	8	-.1336	-.1711	-.1267	-.8544
	9	-.3569	-.4086	-.1049	-2.0221
	10	.0063	.0633	.3385	.8388
	11	-.0146	-.1470	-.1859	-1.9475
E	12	1.0000	.8621	.3824	3.9435
	13	-.2774	-.3554	-.2631	-1.7745
	14	-.2523	-.2888	-.0741	-1.4293
	15	-.0001	-.0009	-.0049	-.0122
	16	.4702	.2170	.0403	.7275
	17	-1.0000	-.4615	-.0857	-1.5472
X	18		1.0000	.5271	4.5098
	19		-.4552	-.3370	-2.2731
	20		-.3306	-.0849	-1.6360
	21		-.0093	-.0498	-.1233
	22		-.1001	-.0186	-.3357
	23		.1048	.0368	.1417
	24		-1.0000	-.3511	-1.3521
Reverse	1		.3511	.3511	
	2	-.0763	-.1620	.0857	

Continuing as above, successively interchanging columns and blocks C and D, B and C and finally A and B from their position in Table 11 and doing the increasingly greater amount of calculation each time, we obtain successively $\beta_{CX} = .2041$, $\beta_{BX} = .6399$ and $\beta_{AX} = .9822$. From these and their previously calculated conjugates we obtain

$$r_{CX \cdot ABDE} = .1253, \quad r_{BX \cdot ACDE} = .4436, \quad r_{AX \cdot BCDE} = .4936$$

The amount of additional work is not great, and the information obtained may be of great importance.

We reach the conclusion in our illustrative problem that land value (X) is more highly correlated with percentage of land in corn (E), independent of the associations of E and X with the other four variables, than it is with any of the other variables which we have considered. This is the same conclusion, though with somewhat different quantitative relations, that was deduced from the partial regression coefficients and the regression equation worked out in Part III.

If the investigator does not care for the simple (zero order) correlation coefficients (Table 7), but wishes only the highest order partial correlation coefficients, together with the multiple regression equation, he may avoid the calculation of the r's and proceed directly to the solution of the normal equations using only the product-moments as the necessary data. See Tolley and Ezekiel, "A Method of Handling Multiple Correlation Problems", Quar. Pub. Am. Statistical Asso., Dec. 1923.

We shall close this part by returning to the problem first considered; namely, the correlation between number of brood sows per 1,000 acres (D) and percentage of land in corn (E); but now we shall find what it would be independent of *all* the other variables. The symbol is $r_{DE \cdot ABCX}$.

This may be found by continuing the process first described, calculating in all ten partial correlation coefficients of first order (such as $r_{DE \cdot X}$), six of second order (such as $r_{DE \cdot AX}$), three of the third order (such as $r_{DE \cdot ABX}$) and finally the one required.

The alternative is to calculate β_{DE} and β_{ED} as in Part III and take the square root of their product. This is very easily done. Simply return to Tables 10 and 11 and compute the reverse as far as line 2 in each table. Then read from reverse line 2, column D, in Table 10 the result, $\beta_{ED} = -.0415$, and in Table 11 $\beta_{DE} = -.0763$. It is obvious that the amount of new calculation is trivial. Not only have we obtained the required β's, but we also have an illustration of the important fact that *two conjugate regression coefficients such as these must always have the same sign,* either both positive or both negative. Furthermore, *the corresponding partial correlation coefficient takes the same sign as the two β's have.* Hence, finally

$$r_{DE \cdot ABCX} = -\sqrt{(-.0763)\ (-.0415)}$$
$$= -.06$$

The reader who is interested in partial correlation will find an excellent discussion by Mordecai Ezekiel in the Journal of Farm Economics, Vol. 5, No. 4, pp. 198-203, "The Use of Partial Correlation in the Analysis of Farm Management Data".

An interesting case of partial correlation is that in which time enters as one of the variables. In such "time series", there is frequently a high correlation merely because two variables are changing with time in some regular way, though they may have no conceivable relation to each other. Also, two variables may have a certain periodicity which affects their correlation. For a discussion of this, see H. L. Rietz, "Handbook of Mathematical Statistics", Chapter X.

PART V. CODING

Coding, in the method herein presented of preserving the identity of the individual observations, is the equivalent of the usual grouping into classes and translating the origin of meas-

TABLE 12. CODED VALUES

County	A	B	C	D	E	X	S
	Yield -30	No Change	Acres $\div 5$	Sows $\frac{-10}{2}$	% $\frac{-10}{2}$	\$ $\div 10$	
Allamakee	10	11	21	11	2	9	64
Bremer	6	13	20	19	10	13	81
Butler	4	19	27	16	10	17	93
Calhoun	11	33	32	15	14	28	133
Carroll	9	25	31	27	12	26	130
Cherokee	12	23	33	32	12	27	139
Dallas	10	22	26	16	14	24	112
Davis	1	9	24	0	5	10	49
Fayette	6	13	21	16	8	14	78
Fremont	4	17	27	20	15	21	104
Howard	0	18	27	10	4	12	71
Ida	10	23	37	38	10	27	145
Jefferson	7	14	20	10	8	16	75
Johnson	11	13	24	30	9	19	106
Kossuth	8	24	35	16	10	20	113
Lyon	8	31	36	26	12	28	141
Madison	4	16	25	12	8	18	83
Marshall	15	19	28	20	12	24	118
Monona	4	20	30	16	10	16	96
Pocahontas	10	30	33	15	14	26	128
Polk	11	22	19	10	12	25	99
Story	12	21	26	17	16	28	120
Wapello	5	16	19	12	6	17	75
Warren	3	18	24	9	7	17	78
Winneshiek	6	18	23	20	6	12	85

urement. The method is explained below, using as illustration the same data as hereinbefore. The authors wish to express their opinion, however, that except for purposes of illustration, coding is not desirable in handling so small a number of observations. The time taken in coding more than offsets the time saved in calculation.

Coding is desirable first, if the individual observations on one or more of the variables are numbers of more than two digits; and second, if the number of observations is so large as to warrant the use of punched cards with a sorting machine. It should be distinctly understood that the only purpose of coding is to save time.

The coded values of our variables are given in Table 12. The values of the coded A-variable are formed by subtracting 30 bu. per acre (the smallest yield in the original list) from each original observation. No change is made in the values of B. Each original value of C is divided by 5 and the numbers "rounded" in the usual manner. The original D-values are first divided by 2, then decreased by 10; while the E-values are first decreased by 10, and the results divided by 2. The X-values are divided by 10.

We wish to emphasize the fact that for purposes of illustrating the various ways of coding, we have greatly overdone the thing. Ordinarily subtraction should be confined to such easily subtracted numbers as 10, 50, 100, etc. while division should be limited to division by 10, 100, etc.

It is desirable that coded values should all be less than 100, and usually they should be less than 50. Coding by addition can be used to eliminate negative values of a variable, and coding by multiplication can be used to eliminate decimals.

Coding by subtraction (or addition) does not affect the standard deviations or the correlation coefficients at all. A coded mean, however, is less (or greater) than the true mean by exactly the same amount as the coded value of an observation is less (or greater) than its true value. Thus, Table 13 shows that $\sigma_A = 3.70$ bu. per acre, just as it is if calculated from true values but the coded mean of 7.48 must be increased by 30 to equal the true mean of 37.48 bu. per acre.

Coding by division (or multiplication) has no effect on correlation coefficients, but produces a corresponding division (or multiplication) of both the mean and the standard deviation. If, however, division is accompanied by a "rounding" of the resulting coded values, as is practically always the case, small discrepancies will exist between the true means and standard deviations and those obtained by adjusting the coded means and standard deviations. However, if the coding is not radical, the differences are always small compared to their standard devia-

TABLE B CODED VALUES CALCULATION OF CORRELATION COEFFICIENTS

	A	B	C	D	E	X	S
Sums	187	488	668	433	246	494	2516
Means	7.48	19.52	26.72	17.32	9.84	19.76	100.64
A_1	1741	3879	5121	3615	2013	4070	20439
A_2	1399	3650	4997	3239	1840	3695	18820
A_3	342	229	124	376	173	375	1619
A_4	18.49	552.3	493.7	749.4	323.9	559.9	
B_1		10418	13635	8885	5141	10375	52333
B_2		9526	13039	8452	4802	9643	49112
B_3		892	596	433	339	732	3221
B_4		29.87	797.5	1210.6	523.3	904.5	
C_1			18562	12171	6783	13703	69975
C_2			17849	11570	6573	13200	67228
C_3			713	601	210	503	2747
C_4			26.70	1082.2	467.8	808.5	
D_1				9143	4537	9221	47572
D_2				7500	4261	8556	43577
D_3				1643	276	665	3995
D_4				40.53	710.1	1227.2	
E_1					2728	5309	26511
E_2					2421	4861	24757
F_3					307	448	1754
E_4					17.52	530.5	
X_1						10678	53356
X_2						9761	49716
X_3						917	3640
X_4						30.28	
σ	3.70	5.97	5.34	8.11	3.50	6.06	
		Adjusted Values					
Means	37.48	19.52	133.60	54.64	29.68	197.60	
σ's	3.70	5.97	26.70	16.22	7.00	60.60	

tions, and are not, therefore, statistically significant. For example, coded $\sigma_C = 5.34$, giving an adjusted $\sigma_C = 5.34 \times 5 = 26.70$ acres, which should be compared with the true $\sigma_C = 26.78 \pm 2.56$. Coded $M_x = 19.76$, adjusted $M_x = 10 \times 19.76 = \$197.60$ per acre, and true $M_x = \$198.20 \pm \$8.26$. (Sheppard's corrections may be applied. Kelley: "Statistical Method", p. 167.)

In cases where both subtraction and division have been used in coding, means must be adjusted by both multiplication and addition, the order being the reverse of that used in coding. Standard deviations must be adjusted by multiplication only. As a first example, consider variable D. In coding, values of D were first divided by 2 and then 10 was subtracted from the quotient. In adjusting the coded standard deviation, we merely multiply by 2 but in adjusting the coded mean, we must first add 10, then multiply by 2. (See Table 13.) As a second example, E was coded by first subtracting 10 and then dividing by 2. To adjust the mean, therefore, we must first multiply the coded mean by 2 and then add 10 to the product. To adjust the standard deviation, simply multiply by 2.

Means and standard deviations must be adjusted before use in the regression equation.

The values of coded S are found by adding the several coded values of the other variables without any reference whatever to the values of S used in Table 6.

Table 14 exhibits the simple correlation coefficients obtained by using coded values. In no case is there a significant divergence from the values given before.

TABLE 14. CORRELATION COEFFICIENTS FROM ORIGINAL DATA AND FROM CODED DATA

	B		C		D		E		X	
	Original	Coded	Original	Coded	Original	Coded	Original	Coded	Original	Coded
A	.41	.42	.25	.25	.50	.50	.53	.53	.67	.67
B			.75	.75	.34	.36	.68	.65	.80	.81
C					.57	.56	.50	.45	.64	.62
D							.38	.39	.53	.54
E									.86	.84

The value of R calculated from these data is .94. The regression coefficients are contrasted in Table 15.

The differences are comparatively great, but the statistical significance of the results is little altered, even though the coding was intentionally carried to excess.

TABLE 15. REGRESSION COEFFICIENTS

Partial Regression Coefficient of X on	A	B	C	D	E
β's calculated from original data	.25	.31	.08	.08	.45
β's calculated from coded data	.22	.38	.03	.12	.42

PART VI. PRECAUTIONS AND SUGGESTIONS

Before any correlation study is undertaken, it is important to make a serious effort to think through the nature of the causes connecting the variables. Much valuable time and effort are wasted by rushing into elaborate calculations before a definite plan is formulated. Many students, laboriously working out correlation coefficients, feel that their work must have a certain virtue simply because they have spent so much time in calculation. On the other hand, preliminary correlation studies are often indispensable as a guide to the formulation of the final plans even though the latter may not include the correlation methods at all.

Cause and effect cannot be determined by correlation. Two variables may be constantly and intimately associated and yet have no causal relations whatever. The correlation coefficients merely point the way to further study and investigation.

Utter familiarity with the data is a prerequisite to successful deductions. Correlation is not a magic formula. Mere calculation, no matter how intricate or extensive, can never take the place of intimate, "common sense" knowledge of the records. Only the man who has worked over his material from many angles until he has become thoroughly familiar with it can hope to apply correlation coefficients and regression lines in a truly fruitful way.

There is a tendency to look upon correlation coefficients as an end in themselves. In some cases, the mechanical labor absorbs so much energy and time that there is very little left for the real job of interpretation. In reality, the correlation coefficients and related constants are usually just a beginning in any serious study. Unless hard thinking and common sense are used in interpretation, correlation work may do more harm than good.

Two extremes should be avoided in your attitude toward the correlation results. On the one hand do not be discouraged if the correlation coefficient is lower than expected, or if the estimated values of the criterion vary widely from actual. Study with the greatest care the cases which deviate most widely. Are they due to accidental or unusual circumstances, and can such

be avoided? Should the relationship be expressed by a curved regression line rather than by one which is straight? Is it necessary to include other variables to account for the discrepancies? Remember, it is not impossible that important discoveries can be initiated by first learning that expected correlation does not really exist. On the other hand, do not be too easily satisfied. It would be a shortsighted policy to stop with a correlation coefficient of .96 when a more perfect explanation might be readily apparent after a little further work.

If the number of independent variables is large and the number of observations relatively small, the multiple correlation coefficient seems to gather a certain amount of "fictitious correlation" merely from the multiplication of the number of variables. B. B. Smith has a correction formula to be used in such cases. This is expected to appear in the March, 1925, issue of the Journal of the American Statistical Society.

The formula is

$$(\text{Corrected } R^2) = 1 - \frac{1 - R^2}{1 - \frac{M}{N}}$$

where M is the number of independent variables and N is the number of observations.

What is the real object of correlation coefficients and their related concepts? The details vary with the field of investigation, with the particular problem in hand, and with the mental peculiarities of the investigator. The purely scientific effort to determine causal relations, the prediction of market prices, vocational guidance, educational policies, the correct method of scoring corn, heredity, land values, the correction of yields for soil variation,—these are some of the problems attacked with correlation methods. The research worker must always interpret his results in the light of his own knowledge. After all, correlation is simply one scheme for discovering and evaluating relationship.